Databases and Data Mining

Carolyn J. Lawrence and Doreen Ware

Abstract Over the course of the past decade, the breadth of information that is made available through online resources for plant biology has increased astronomically, as have the interconnectedness among databases, online tools, and methods of data acquisition and analysis. For maize researchers, the number of resources available is both impressive and daunting, in many cases leaving them at a loss regarding where to begin. Described here is an historical perspective on the origin of these resources, as well as how they are expected to change and grow in the future. We outline the current types of resources, how they are connected, and methods for data acquisition, analysis, and interpretation. In addition, we offer guidance to assist researchers place data generated by their maize projects into appropriate databases for long-term storage and use.

1 Databases Past and Present

The theory for storing information in relational databases was reported in 1970 by Edgar Codd, who worked for IBM Research (Codd 1970). Subsequently, various methods for storing data relationally were implemented based upon Codd's ideas. Early on, these data resources could only be accessed by direct interaction with the computers that stored the data. However, the creation of ARPANET (the U.S. government's Advanced Research Projects Agency's networking project) in the early 1970s served as the basis for linking various resources together. ARPANET eventually

C.J. Lawrence
USDA-ARS, Corn Insects and Crop Genetics Research Unit and Iowa State University, Departments of Agronomy and Genetics, Development and Cell Biology
carolyn.lawrence@ars.usda.gov

D. Ware
USDA-ARS, U.S. Plant, Soil and Nutrition Research Unit
doreen.ware@ars.usda.gov
Cold Spring Harbor Laboratory
ware@cshl.edu

J.L. Bennetzen and S. Hake (eds.), *Maize Handbook - Volume II: Genetics and Genomics*,

evolved into the present day Internet, which has brought the utility of databasing to bear on problems ranging from personnel management to shopping online.

Simultaneous with the evolution of database technologies and the creation of the Internet, biologists began to create datasets of ever-increasing size. These datasets included DNA sequence information and molecular biological data, as well as others that were species-specific in nature. A need to store, categorize, and easily access these datasets resulted in the adoption of database technologies by biologists. Coupled with tool-building activities for biological data analysis, the field of bioinformatics was born.

Some of the earliest and most widely utilized publicly accessible biological databases were created to store DNA sequences and to make the sequences accessible via a variety of methods. These include EMBL (the European Molecular Biology Laboratory), DDBJ (the DNA Data Bank of Japan), and GenBank. The first of these, EMBL, which is run by the European Bioinformatics Institute (EBI), began in 1980 (Stoesser et al., 1997), whereas DDBJ (http://www.ddbj.nig.ac.jp/) began work in 1986 (Tateno and Gojobori 1997), and NCBI (the National Center for Biotechnology Information) founded GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) in 1992 (Benson et al., 1997). All are permanently funded, long-term repositories. To ensure that each of these three equivalent repositories could serve the most comprehensive and up-to-date set of sequences, each agreed to share their data with the other two when all became part of the International Nucleotide Sequence Databases Collaboration.

The plant biology databases AAtDB (An *Arabidopsis thaliana* Database, which later evolved into the *Arabidopsis thaliana* Database, AtDB, then The *Arabidopsis* Information Resource, TAIR; Flanders et al., 1998; Huala et al., 2001) was one of the first plant biological databases to be created. Howard Goodman founded AAtDB in 1991 as a resource to serve information on the model dicot *Arabidopsis thaliana*. Its evolution to become TAIR involved the adoption of AIMS, the *Arabidopsis* Information Management System, which served as the primary stock information and ordering facility of the *Arabidopsis* Biological Resource Center (ABRC). Other plant biological databases that began in 1991 include GrainGenes for the Triticeae (Carollo et al., 2005), RiceGenes for rice (Cartinhour 1997), SoyBase for soy (Grant and Shoemaker, 2007), Dendrome for forest trees (Neale, 2007), and MaizeDB for corn (Polacco and Coe, 1999). MaizeDB, the maize equivalent to AAtDB, was headed by USDA-ARS scientist and past editor of the Maize Newsletter, Ed Coe. MaizeDB served genetics information including (but not limited to) maps, phenotypes, and molecular marker/probe data. The current maize database, MaizeGDB, came into existence when MaizeDB merged with a sequence database called ZmDB (Lawrence et al., 2004). Like AtDB/TAIR, MaizeDB/MaizeGDB also stores data for the Maize Genetics Cooperation–Stock Center (Scholl et al., 2003). More recently, Gramene, a resource for comparative biology among grass species, was established (Ware et al., 2002), and various maize project-specific resources have come on line. GrainGenes, SoyBase, and MaizeGDB operate on permanent funds from the USDA-ARS. Dendrome is permanently funded by the U.S. Forest Service, and the others are not funded long-term.

1.1 Types of Resources

Databases storing genomic information fall into various categories based upon their role within a larger context. A Laboratory Information Management System (LIMS) is the most basic sort of database and interface solution, and can be as simple as a spreadsheet stored on a computer in a particular laboratory. Complex systems where the LIMS is made up of various data pipelines and/or laboratories are generally highly customized and are created to support an individual research group's shared data management needs. Data stored within a LIMS environment represent the group's working information and generally are not made available for use outside of the group that generated the data. In some cases, the LIMS system may eventually be deployed as a public repository, but often with limited support for long-term maintenance. Static Repositories (SRs) are those resources where data (often limited to a single data type) are deposited for long-term storage. The data generally are not changed over time, hence the moniker 'static'. The most well known SR for biological data is GenBank (http://www.ncbi.nlm.nih.gov/Genbank/; Benson et al., 2007), the federally funded resource that stores sequence data for all species. An Automatic Annotation Shop (AAS) harvests data from SRs and runs those data through analysis pipelines to create products that have added value for use by researchers. Of these, JCVI (the J Craig Venter Institute, formerly TIGR, http://www.jcvi.org) is an AAS that provides value-added sequence-based products, including genome assemblies and repeat databases based upon the sequence set stored at GenBank (Chan et al., 2006). Model Organism Databases (MODs), which are generally species-specific, have been created for various plant species, including soybean (SoyBase; http://www.soybase.agron.iastate.edu), *Arabidopsis* (The *Arabidopsis* Information Resource; http://www.arabidopsis.org/), and various other species including *Drosophila*, *C. elegans*, mouse, zebrafish, etc. (Crosby et al., 2007; Bieri et al., 2007; Eppig et al., 2007; Sprague et al., 2006). These databases are built and maintained by teams of information technology specialists and biological curators, and represent highly curated products that recapitulate the biology of a particular species by storing species-specific data types and making available specialized tools for analyzing those data within their specialized biological context. Most MODs store and integrate more than one data type and provide the community with integrated views and specialized tools for analyzing those data within the context of their organism of interest. Clade-Oriented Databases (CODs) store and make accessible those data that can be leveraged by researchers to enable comparative biological analyses, including sequence similarity and genomic synteny information. The CODs are especially important for communities working on groups of species simultaneously, such as potato, tomato, and pepper (SGN; http://www.sgn.cornell.edu; Mueller, 2005). Other CODs include LIS (the Legume Information System; Gonzales et al., 2005) and GrainGenes (for small grains; Carollo et al., 2005).

MaizeGDB (http://www.maizegdb.org/; Lawrence et al., 2007) is the MOD for maize. It is the central repository for all sorts of maize genetics and genomics data, and includes information on maps, loci, gene products, molecular markers,

and references, as well as bulletin boards, such as a maize-specific job list and a calendar of upcoming events. MaizeGDB also serves as the clearinghouse for maize nomenclature and supports the activities of the Maize Nomenclature Committee (http://www.maizegdb.org/maize_nomenclature.php) and the Maize Genetics Executive Committee (http://www.maizegdb.org/mgec.php). To best determine how to move MaizeGDB forward to meet the needs of the maize community, a Working Group meets yearly (current membership is listed on the home page at http://www.maizegdb.org), and feedback from researchers who utilize MaizeGDB is solicited, both through the Web interface and in person at meetings, including the Annual Maize Genetics Conference and the International Plant and Animal Genome Conference. Sets of data are taken in over the course of the year by data type (see http://www.maizegdb.org/data_schedule.php), and methods for collaborating with MaizeGDB personnel to incorporate researchers' data into the database are also available online (see http://www.maizegdb.org/data_contribution.php).

Gramene (http://www.gramene.org/; Jaiswal et al., 2006) is the COD that serves maize data alongside information from other grasses to enable cross-species comparisons, including the analysis of synteny information among cereals, which is useful for leveraging data from other grasses to advance maize research.

Other maize resources currently in operation include Panzea (http://www.panzea.org/; Zhao et al., 2006), the Maize WebFPC (http://www.genome.arizona.edu/fpc/maize/; Gardiner et al., 2004), the Maize Genome Sequencing Consortium's genome browser (MaizeSequence.org; http://www.maizesequence.org/), PlantGDB's maize genome browser, which is called ZmGDB (http://www.plantgdb.org/ZmGDB/; Schlueter et al., 2006), the Functional Genomics of Maize Chromatin Consortium database (http://www.chromatin-consortium.org/), and MAGI (Maize Assembled Genomic Island; http://www.magi.plantgenomics.iastate.edu/; Fu et al., 2005). A non-exhaustive list of additional plant-specific databases that are used by maize researchers is shown in Table 1.

Table 1 Online resources utilized by maize researchers that are not maize-specific.

Resource Name	Resource Type	Link	Funding Source(s)
ChromDB	LIMS	http://www.chromdb.org/	NSF
GrainGenes	COD	http://wheat.pw.usda.gov/	USDA-ARS
Gramene	COD	http://www.gramene.org/	NSF, USDA-ARS
GRIN	Static/LIMS	http://www.ars-grin.gov/	USDA-ARS
NCBI (esp. PLANTS)	Static	http://www.ncbi.nlm.nih.gov/ and http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html	NIH
PlantGDB	COD/AAS/LIMS	http://www.plantgdb.org/	NSF
PLEXdb	AAS	http://www.plexdb.org/	NSF, USDA-ARS, USDA-CSREES
TAIR	MOD	http://www.arabidopsis.org/	NSF
JCVI	AAS	http://www.jcvi.org/	NSF
UniProt	Static	http://www.pir.uniprot.org/	NIH

1.2 *Interconnections among Different Repository Types*

Online resources abound for maize. This creates an environment that both assists and stymies researchers. Because many resources are available, it is probable that some available resource stores the data that could help to address a given research question. However, the breadth of resources, coupled with few or no mechanisms to search all resources simultaneously, makes exhaustive searches of available data difficult, if not impossible. Methods that currently are used to interconnect repositories are outlined below.

At the most basic and straightforward level, online resources are interconnected using Web-based hypertext links. Links are stored in a data repository and can provide context-sensitive points of entry into relevant data hosted at another site. Three other methods for interconnecting data among different repositories based on methods and data architectures include the following: data warehousing, data federation, and the use of mediators or portals (reviewed in Lushbough et al., 2008).

Data warehousing represents the least cost-effective method of interconnecting data repositories (Lacroix and Critchlow 2003). In this type of set up, one database duplicates some of the data from another repository. Data federation requires cooperation among all members of the federation (Sheth and Larson 1990). It consists of component databases that are autonomous yet participate in the federation to allow partial and controlled sharing of their data. A mediator is the most flexible approach to data integration. It offers intermediary services that link the data resources and application programs. Mediator approaches integrate information by accessing and retrieving data from multiple resources, abstracting and transforming the retrieved data, integrating the product, and processing the integrated data to return a result (Wiederhold and Genesereth 1997).

In practice, biological databases use all these approaches to integrate data, although data warehousing is probably the most commonly used method of data integration. The use of federation and mediator approaches is growing, due largely to the availability of Web services (technologies that enable one repository to grab information from another resource using defined protocols), ontologies (hierarchical controlled vocabularies that can be used to interconnect related information) and other controlled metadata tags, and the semantic web (a standard for inferring the meanings associated with shared data).

Interconnections via links are often supported by minimal data warehousing. For example, by warehousing GenBank identifiers as well as a protocol for linking to GenBank, any repository could embed linkages from, e.g., molecular markers available at a resource to the marker's GenBank sequence record. A more significant instance of data warehousing is the inclusion of maize molecular markers at Gramene: The marker data were contributed for inclusion in Gramene from MaizeGDB and are currently represented at both repositories.

The NCBI (National Center for Biotechnology Information; Wheeler et al., 2005) is comprised of various databases, including PubMed, GenBank, and various specialized resources, such as PLANTS (http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html). These resources represent a database federation, in that they are each

separate databases, but are presented as if they were components of a single repository (see http://www.ncbi.nlm.nih.gov/Database/). Gramene also uses a similar strategy: the genome information, diversity data, pathway information, map, and protein data are housed in separate database structures, but are presented as if they were a single repository within Gramene. In the case of the Genome Browser, Gramene and the maize sequencing project make use of Ensembl (Fernández-Suárez and Schuster, 2007) to store and visualize the data. Currently there are seven sequenced genomes available, 4 monocots, two varieties of rice, maize, sorghum, grape, poplar and sorghum, and 3 dicots, including Arabidopsis. For diversity data, Gramene leverages both the Ensembl and Genomic Diversity and Phenotype Data Model GDPDM to store and distribute diversity data. Currently, Gramene hosts diversity data for maize, rice, and wheat. In the case of pathways, data are stored and displayed using the SRI pathway tools (Hubbard 2002). Gramene supports Web services through Ensembl and GDPC. Ensembl uses the distributed annotation services DAS architecture (Dowel 2001).

A more recent example of implementation of Web services within the plant community is VPIN (the Virtual Plant Information Network; http://vpin.ncgr.org/), which makes use of a mediator approach for data sharing and presentation. VPIN serves data from Gramene, JCVI, and other databases. The underlying technology is SSWAP (Simple Semantic Web Architecture and Protocol), which was developed by D. Gessler and others. Using SSWAP, resources are allowed to define themselves on the Web. Defined documents are available via a non-exclusive discovery server (http://sswap.info) and also can be accessed by third-party servers. Classes of data are deduced based on shared properties, or finding resources based on the type of data they accept (instead of the resource's static categorization). Shown below are some example implementations for each method for creating these interconnections.

MaizeGDB currently uses link integration and data warehousing to enable researchers to get to data of interest stored at sites other than http://www.maizegdb.org/. Linkages from MaizeGDB include (but are not limited to) the following: BioCyc, CerealsDB, Dana-Farber Cancer Institute, Gramene, KEGG, the Maize Genetics Cooperation–Stock Center, MaizeSequence.org, NCBI, PlantGDB, SwissProt/

Steps	Data-Sharing Approach	Example Link(s)
Jump to relevant ZmGI contigs at the Dana-Farber Cancer Insitute from a sequence page at MaizeGDB	Link Integration	http://www.maizegdb.org/cgi-bin/displayseqrecord.cgi?id=BG836376
Find out to which contig at MaizeSequence.org the locus *bnlg1372* belongs via MaizeGDB.	Data Warehousing	http://www.maizegdb.org/cgi-bin/displaylocusrecord.cgi?id=144892
Jump from PubMed's display of Wang and Dooner, 2006 to nucleotide sequences and taxonomy via "Links"	Federation	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=ShowDetailView&TermToSearch=17101975
Find out (via SSWAP at VPIN) how to create direct linkages to QTL trait symbols at Gramene	Mediation	http://www.sswap.org/ with query "qtl trait symbol"

TREMBL, JCVI, WebFPC, and ZmGDB. Project-specific databases to which MaizeGDB links include ChromDB, the Chromatin Consortium, the Maize TILLING Project site, MAGI, MaGMaP, and others. A more complete list of current maize projects can be accessed at MaizeGDB on the Maize Research Projects page (http://www.maizegdb.org/maizeprojects.php/). Similarly, Gramene uses a combination of linked integration, data warehousing, and Web services methods for obtaining data internally (using a federated approach) and externally, as described above.

MaizeGDB is the hub or focal point for connecting a researcher to relevant resources when initiating a search from a maize-centric perspective. A researcher can, for instance, navigate to GenBank to access the sequences of relevant BACs, or navigate to Gramene to help identify orthologs in rice, wheat, or other grasses. Gramene provides a central location for rice and plant researchers when initiating searches related to rice biology or for cross-species analysis for plants.

2 Data Mining using Currently Available Resources

Databases are only as useful as the information they can provide for given research questions. Below is an example problem that a researcher could solve online using various resources. In addition to access through Web-based displays, in many cases resources offer access to the database through an application programming interface (API) or via wholesale downloads of the database.

2.1 Example Problem 1: Discovering and Developing Molecular Markers for a Genomic Region of Interest Given some Sequence Data

Researcher 1 has created a recessive mutation that disrupts meiotic spindles using Robertson's *Mutator* (*Mu*) (reviewed in Lisch et al., 1995; Lisch, this volume). She also has found that the marker for *bnlg1185* on chromosome 10 identifies a locus within ten centiMorgans (cM) of the mutation responsible for the mutant phenotype. In an effort to narrow down the region before she tries walking to the gene, she plans to check to see which markers might lie closer to the mutation and will find out whether the Maize Genome Sequencing Consortium has sequenced a BAC containing the markers identified.

First she goes to MaizeGDB and uses the search field at the top of any MaizeGDB page (Figure 1A) to search loci for 'bnlg1185'. The locus *bnlg1185* (http://www.maizegdb.org/cgi-bin/displaylocusrecord.cgi?id=144839) is at the top of the results list. On that page, she clicks the link to see the IBM2 2004 Neighbors 10 map. On that page, the locus *bnlg1185* is highlighted in green (Figure 1B). Nearby loci that could be tested to orient the mutation's position relative to *bnlg1185* are *gln1*, which is 20.19 cM proximal to *bnlg1185*, and *csu48*, which is 27.69 cM distal to *bnlg1185*.

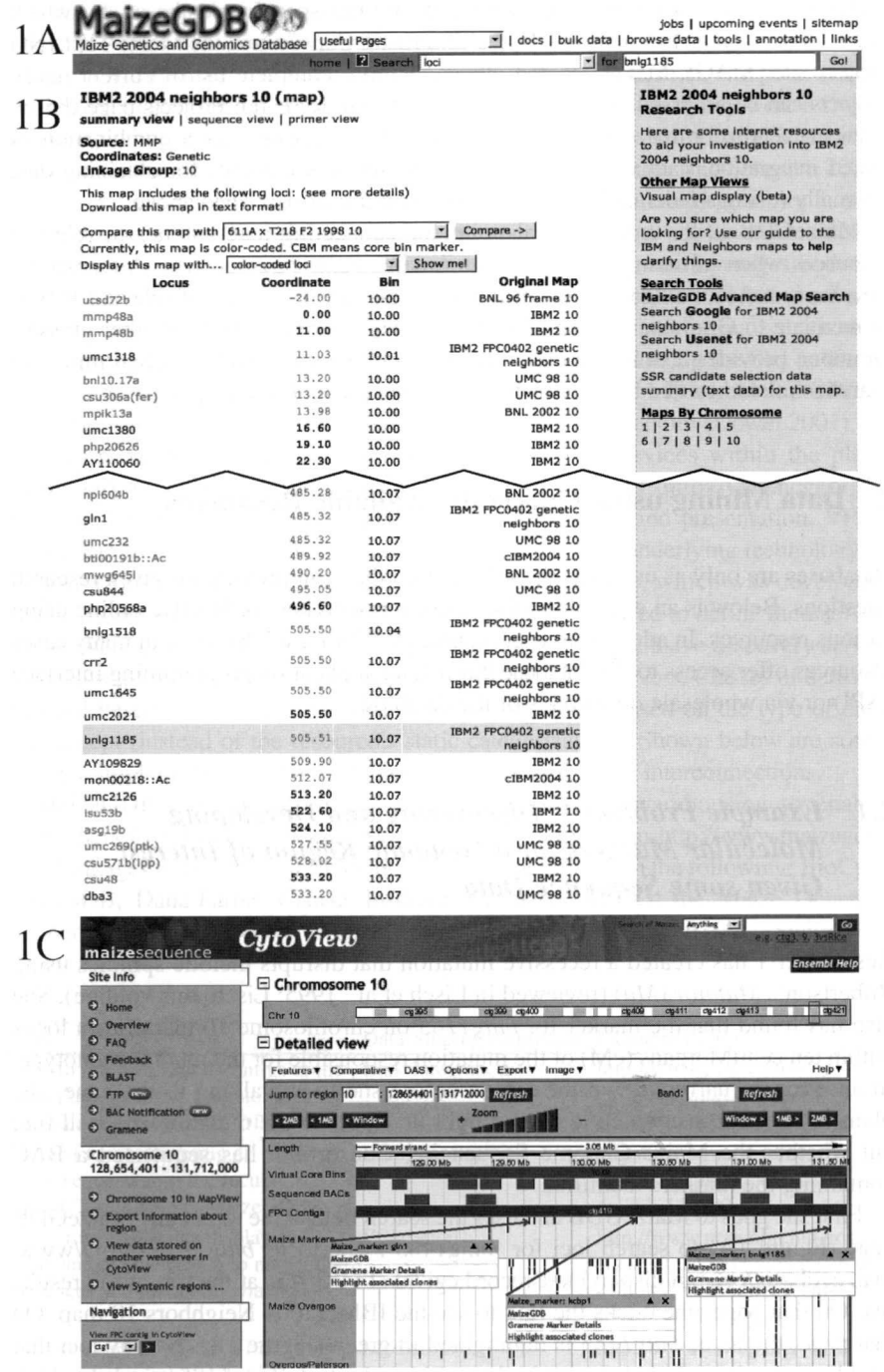

Fig. 1 Screen shots of the MaizeGDB and MaizeSequence.org displays used by Researcher 1. 1A shows the search field at the top of any MaizeGDB page, which can be used to find the locus

Using these two markers, she can orient her mutation via recombination mapping. It turns out that her mutation is approximately 10 cM from *bnlg1185* and 40 cM from *csu48*. This indicates that the mutation lies between *gln1* and *bnlg1185*.

Next she wants to narrow the interval containing the mutation of interest by selecting markers between *gln1* and *bnlg1185*. She, once again, navigates to MaizeGDB and uses the search field at the top of any page, this time using 'gln1' as a search term. Toward the top of the *gln1* locus page (http://www.maizegdb.org/cgi-bin/displaylocusrecord.cgi?id=61733) is a note that "This locus is part of contig ctg419 at MaizeSequence.org." Similarly, the locus page at MaizeGDB for *bnlg1185* says that the locus is associated with "contig ctg419". Clicking the link to view that contig at MaizeSequence.org, she winds up at the 'CytoView' display of the contig (Figure 1C). Mousing over the row of Maize Markers beneath ctg419, she finds vertical bars indicating the relative locations of *gln1* and *bnlg1185*. Clicking on nearby markers, she sees links to information at MaizeGDB and Gramene, as well as a tool that enables her to highlight associated clones of interest.

Interestingly, on the 'CytoView' display at MaizeSequence.org, a marker labeled 'kcbp1' lies between her markers of interest. When she visits the MaizeGDB locus record for *kcbp1*, she finds that it is, "expressed in all tissues with highest levels in actively dividing cells." Encouraged that this might in fact be (or be very similar to) her gene of interest, she develops primers that are specific for *Mu* as well as primers that are specific to *kcbp1*. If this is the gene where *Mu* inserted, she should be able to get a product with PCR. If not, it's back to narrowing the region by mapping new markers between *bnlg1185* and *gln1*, and subsequently trying out combinations of gene-specific and *Mu*-specific primers with PCR.

2.3 What to do if no Resources Support Your Workflow Needs

In instances where the data exist at a particular repository that can be pieced together to answer a particular researcher's questions, but the method by which the data are made available causes the same repetitive set of steps to be carried out many times, the researcher's best course of action is to use links at the repository to contact its personnel, describe the special needs of the project at hand, and to ask for the creation of a customized dataset. Before making contact, it is advisable for the researcher to document exactly how s/he can use use the repository's data to get the necessary set of information, such that the repository's personnel could follow that protocol. Furthermore, it is advisable for the researcher to accompany the

Fig. 1 (continued) *bnlg1185*. From the *bnlg1185* locus page, clicking the link to the IBM2 2004 Neighbors 10 map not only shows the map of interest (1B), but also highlights the location of *bnlg1185*. Using links at the top of the *bnlg1185* and/or *gln1* page, the MaizeSequence.org CytoView page for contig ctg419 is accessed. There (1C), the relative positions of *gln1*, *kcbp1*, and *bnlg1185* are shown

request with a table including at least one example of the desired data. It is often the case that these processes can be automated and will be of use to other researchers in the field. In most instances, a researcher who requests a dataset will receive it promptly, thus allowing valuable time to be spent at the bench, rather than clicking through an interface to collect a set of data.

3 How to get Your Own Project Data into the Mainstream Databases

Researchers often begin to generate data and to store generated data well in advance of knowing whether the data will be useful—after all, it's called research, right? One problem that often arises when preliminary investigations go well and data generation picks up is a tendency to continue to store generated data in an idiosyncratic manner. Datasets that are not formatted in a way that integrates well with existing information stored at the larger data repositories pose a real problem for subsequent integration and results in data quality issues, loss of data, and sometimes a loss of human resources available to the project.

3.1 Choosing a Repository to House Data

The best way for a researcher to find a repository to house project data is to first consider which repositories already hold the types of data to be generated. With that list in hand, the researcher should consider how others might be expected to utilize the data and then contact the repository that best meets the needs of that type of data and analysis. Feedback links at the repository of interest or direct contact with the repository's lead scientist or project manager are two avenues of initial inquiry as to whether the data could be accommodated.

Funding sources affect resource development and maintenance practices as well as data access longevity. For this reason, it is always advisable for a researcher to consider whether the resource most appropriate for storing the data has long-term funding, as well as to inquire about whether and how the research project's funds could support data integration directly. Because allocating funds to the repository may be required, it is important to investigate data warehousing options well in advance of writing proposals to funding agencies.

3.2 Data Types that Help with Resource Integration

Once a repository has been selected for collaboration, it is wise for the researcher to depend upon the personnel who work at that resource for guidance in how to store project data in a manner that will facilitate its incorporation into the data repository. Some data types that are likely to be useful for integration are the following:

Controlled vocabularies — These resource-specific sets of words are assigned to records to enable others to find the data via keyword searches. One example would be assigning the keyword "SSR" to a probe/molecular marker record that is to be included in MaizeGDB. Note that, if the researcher were to fail to inquire in advance how the repository assigns terms, it is possible that the word "microsatellite" would have been assigned, thus causing the new records to be absent from SSR lists generated by the Web interface or other data presentation interfaces.

Ontologies — Ontologies are hierarchically-related controlled vocabularies that are a standard utilized by many resources (i.e., they are not database- or resource-specific). Again, inquiring with the resource into which data will flow is key to finding out which ontologies to use for descriptor assignments during data collection. Ontologies that are likely to be suggested include the Plant Ontologies (PO; Jaiswal et al., 2005) and Gene Ontologies (GO; The GO Consortium, 2000). Including these terms enables repositories to warehouse links to other resources using the annotations. For example, at PO one can find genes for *Arabidopsis*, rice, and maize that affect inflorescences, or parts of inflorescences, and link to individual database records at TAIR, Gramene, and MaizeGDB for detail.

Size and color standards — To help others to know, for example, the size of bands on gels, lengths of floral organs, or the color of a kernel phenotype; and to enable evolving software tools to search images given an example image as a query (Shyu et al., 2007), researchers should take care to collect these sorts of data and to work with the chosen repository well in advance of data submission to define size and color standards for all phenotypic measurements.

4 The Future of Plant Databases and Data Mining

Resources for database creation and development continually diminish. The repositories upon which maize researchers depend are affected by the scarcity of funding, making it difficult to continue to serve researchers at the level to which they have become accustomed. Simultaneously, many researchers create project-specific databases without ensuring future accommodation by a long-term repository, making it difficult to integrate generated data with related information once project personnel have moved on to other things. The good news is that these problems are apparent, and the funding agencies are responding. A new National Science Foundation-funded project to create a plant cyberinfrastructure called the iPlant™ Collaborative (see http://

www.iplantcollaborative.org) has begun, with the intention to create community-based resources that build upon existing resources and also to encourage the development of new technologies and methods of collaboration to better organize resources and their interactions. In addition, a new Project Portal for corn (POPcorn; also funded by the NSF and to be implemented by USDA-ARS personnel) is in the planning stages. The POPcorn resource, which will be ancillary to MaizeGDB, will allow researchers to search all maize projects' data simultaneously from one Web portal and will provide tools to allow dataset upload to MaizeGDB at a research project's close.

Maize researchers are at the cusp of a new era: The sequence of B73 will be available at the end of 2008, and the cost to sequence other inbred lines falls each day. Maize was once a genetics-rich but sequence-poor model for research, but this is changing. The stage is set for a renaissance in maize research where the species' strengths shine: researchers will have access to sequenced genomes, excellent genetics, and unparalleled cytogenetics. These tools will allow a better understanding of metabolism, development, and breeding. With these excellent assets emerging, the need for a well-annotated genome, improved repositories to store diverse data, and improved connections among maize informatics resources becomes paramount. It is anticipated that NCBI will represent the available maize genome sequence and related data as they have for other sequenced species. For MaizeGDB, the database and Web interface will evolve based upon available resources coupled with the community's stated objectives as communicated by the Maize Genetics Executive Committee and the MaizeGDB Working Group. Increased requirements for data handling will emerge and be met, and researchers' ability to utilize all available data will improve as the data stored in various places are shared by increased utilization of federation and mediation approaches, as well as other technologies currently under development.

References

Benson, D.A., Boguski, M.S., Lipman, D.J., and Ostell, J. (1997) GenBank. Nucleic Acids Res. 25(1), 1–6.

Benson, D.A. Karsch-Mizrachi, I., Lipman P., Gelbart, W.M., and the FlyBase Consortium. (2007) FlyBase: genomes by the dozen. Nucleic Acids Res. 35(Database issue), D486–D491.

Bieri, T., D. Blasiar, P. Ozersky, I. Antoshechkin, C. Bastiani, P. Canaran, J. Chan, N. Chen, W.J. Chen, P. Davis, T.J. Fiedler, L. Girard, M. Han, T.W. Harris, R. Kishore, R. Lee, S. McKay, H.M. Muller, C. Nakamura, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E.M. Schwarz, W. Spooner, M.A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, R. Durbin, L.D. Stein, P.W. Sternberg, and J. Spieth. 2007. WormBase: new content and better access. *Nucleic Acids Res* 35: D506–510.

Carollo, V., Matthews, D.E., Lazo, G.R., Blake, T.K., Hummel, D.D., Lui, N., Hane, D.L., and Anderson, O.D. (2005) GrainGenes 2.0. An improved resource for the small-grains community. Plant Physiol. 139(2), 643–651.

Cartinhour, SW. (1997) Public informatics resources for rice and other grasses. *Plant Mol Biol* 35(1–2),241–251.

Chan, A., Cheung, F., Lee, D., Zheng, L., Whitelaw, D., Pontaroli, A., Sanmiguel, P., Yuan, Y., Bennetzen, J., Barbazuk, W.B., Quackenbush, J., and Rabinowicz, P.D. (2006) The TIGR Maize Database. Nucleic Acids Res. 34, D771–D776.

Codd, E.F. (1970) A relational model of data for large shared data banks. Communications of the ACM 13(6), 377–387.

Dowell, R.D., R.M. Jokerst, A. Day, S.R. Eddy, and L. Stein. 2001. The distributed annotation system. BMC Bioinformatics 2: 7.

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., and the Mouse Genome Database Group (2007) The mouse genome database (MGD): new features facilitating a model system. Nucleic Acids Res. 35(Database issue), D630–D637.

Fernández-Suárez, X.M., and Schuster, M.K. (2007) Using the Ensembl genome server to browse genomic sequence data. Curr Protoc Bioinformatics. 1,1.15.

Fu, Y., Emrich, S.J., Guo, L., Wen, T.J., Ashlock, D.A., Aluru, S., and Schnable, P.S. (2005) Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. Proc. Natl. Acad. Sci. U.S.A. 102(34), 12282–12287.

Gardiner, J., Schroeder, S., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Morgante, M., Landewe, T., Fengler, K., Useche, F., Hanafey, M., Tingey, S., Chou, H., Wing, R., Soderlund, C., and Coe, E.H. (2004) Anchoring 93,971 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. Plant Physiol. 134,1317–1326.

Gonzales, M.D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., Shoemaker, R., Beavis, W.D., and Waugh, M.E. (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. Nucleic Acids Res. 33(Database issue), D660–D665.

Grant, D. and Shoemaker, R.C. (2007) SoyBase, The USDA-ARS Soybean Genome Database. http://soybase.org.

Huala, E., Dickerman, A.W., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, M., Huang, W., Mueller, L.A., Bhattacharyya, D., Bhaya, D., Sobral, B.W., Beavis, W., Meinke, D.W., Town, C.D., Somerville, C., and Rhee, S.Y. (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucleic Acids Res. 29(1), 102–5.

Hubbard, T., D. Barker, E. Birney, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyras, J. Gilbert, M. Hammond, L. Huminiecki, A. Kasprzyk, H. Lehvaslaiho, P. Lijnzaad, C. Melsopp, E. Mongin, R. Pettett, M. Pocock, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik, and M. Clamp. 2002. The Ensembl genome database project. Nucleic Acids *Res* 30: 38–41.

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E., McCouch, S.R., Pujar, A., Reiser, L., Rhee, S., Sachs, M., Schaeffer, M., et al. (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. Comp. Funct. Genomics 6, 388–406.

Jaiswal, P., Ni, J., Yap, I., Ware, D., Spooner, W., Youens-Clark, K., Ren, L., Liang, C., Zhao, W., Ratnapu, K., Faga, B., Canaran, P., Fogleman, M., Hebbard, C., Avraham, S., Schmidt, S., Casstevens, T.M., Buckler, E.S., Stein, L., and McCouch, S. (2006) Gramene: a bird's eye view of cereal genomes. Nucleic Acids Res. 2006 Jan 1;34(Database issue), D717–D723.

Lacroix, Z. and Critchlow, T. (2003) Bioinformatics: Managing Scientific Data. Morgan Kaufmann Publishers, pp. 21–24.

Lawrence, C.J., Dong, Q., Polacco, M.L., Seigfried, T.E., and Brendel, V. (2004) MaizeGDB, the community database for maize genetics and genomics. Nucleic Acids Res. 32(Database issue), D393–D397.

Lawrence, C.J., Schaeffer, M.L., Seigfried, T.E., Campbell, D.A., and Harper, L.C. (2007) MaizeGDB's new data types, resources and activities. Nucleic Acids Res. 35(Database issue), D895–900.

Lisch, D., Chomet, P., and Freeling, M. (1995) Genetic characterization of the *Mutator* system in maize: behavior and regulation of *Mu* transposons in a minimal line. Genetics 139, 1777–1796.

Lushbough, C., Bergman, M.K., Lawrence, C.J., Jennewein, D., and Brendel, V. (2008) BioExtract Server - an integrated workflow-enabling system to access and analyze heterogenous, distributed biomolecular data. IEEE. ACM Transactions on Computational Biology and Bioinformatics. 11

Sept 2008. IEEE computer Society Digital Library. IEEE Computer Society, 10 November 2008 <http://doi.ieeecomputersociety.org/10.1109/TCBB.2008.98.

Mueller, L.A., Solow, T.G., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C., Wright, M.H., Ahrens, R., Wang, Y., Herbst, E.V., Keyder, E.R., Menda, N., Zamir, D., and Tanksley, S.D. (2005) The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. Plant Physiol. 138(3), 1310–1317.

Neale, D. (2007) Dendrome, The USDA Forest Service's Forest Tree Genome Database. http://dendrome.ucdavis.edu.

Polacco, M. and Coe, E. (1999) MaizeDB: The maize database. In *Bioinformatics Databases and Systems*, Letovsky, S.I., ed. Kluwer Academic Publishers, Boston.

Schlueter, S.D., Wilkerson, M.D., Dong, Q., and Brendel, V. (2006) xGDB: open-source computational infrastructure for the integrated evaluation and analysis of genome features. Genome Biol. 7(11), R111.

Scholl, R., Sachs, M., and Ware, D. (2003) Maintaining collections of mutants for plant functional genomics. In Grotewold, E., ed. Plant Function Genomics, Totowa, NJ Humana Press Vol. 236, pp. 311–326.

Sheth, A.P. and Larson, J.A. (1990) Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys. 22(3), 183–236.

Shyu, C., Green, J.M., Lun, D.P.K., Kazic, T, Schaeffer, M., and Coe, E. (2007) Image analysis for mapping immeasurable phenotypes in maize. IEEE Signal Processing Maga. May, 115–118.

Sprague, J., Bayraktaroglu, L., Clements, D., Conlin, T., Fashena, D., Frazer, K., Haendel, M., Howe, D.G., Mani, P., Ramachandran, S., Schaper, K., Segerdell, E., Song, P., Sprunger, B., Taylor, S., Van Slyke, C.E., and Westerfield, M. (2006) The Zebrafish Information Network: the zebrafish model organism database. Nucleic Acids Res. 34(Database issue), D581–D585.

Stoesser, G., Sterk, P., Tuli, M.A., Stoehr, P.J., and Cameron, G.N. (1997) The EMBL nucleotide sequence database. Nucleic Acids Res. 25(1), 7–14.

Tateno, Y. and Gojobori, T. (1997) DNA Data Bank of Japan in the age of information biology. Nucleic Acids Res. 25(1), 14–17.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. Nature Genet. 25, 25–29.

Wang, Q. and Dooner, H.K. (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. Proc. Natl. Acad. Sci. U.S.A. 2006 103(47), 17644–9.

Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., Schmidt, S., Zhao, W., Cartinhour, S., McCouch, S., and Stein, L. (2002) Gramene: a resource for comparative grass genomics. Nucleic Acids Res. 30(Database issue), 103–105.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L., and Yaschenko, E. (2005) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 33(Database issue), D39–D45.

Wiederhold, G. and Genesereth, M. (1997) The conceptual basis for mediation services. IEEE Expert, 12(5), 38–47.

Zhao, W., Canaran, P., Jurkuta, R., Fulton, T., Glaubitz, J., Buckler, E., Doebley, J., Gaut, B., Goodman, M., Holland, J., Kresovich, S., McMullen, M., Stein, L., and Ware, D. (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res. 34(Database issue), D752–D757.